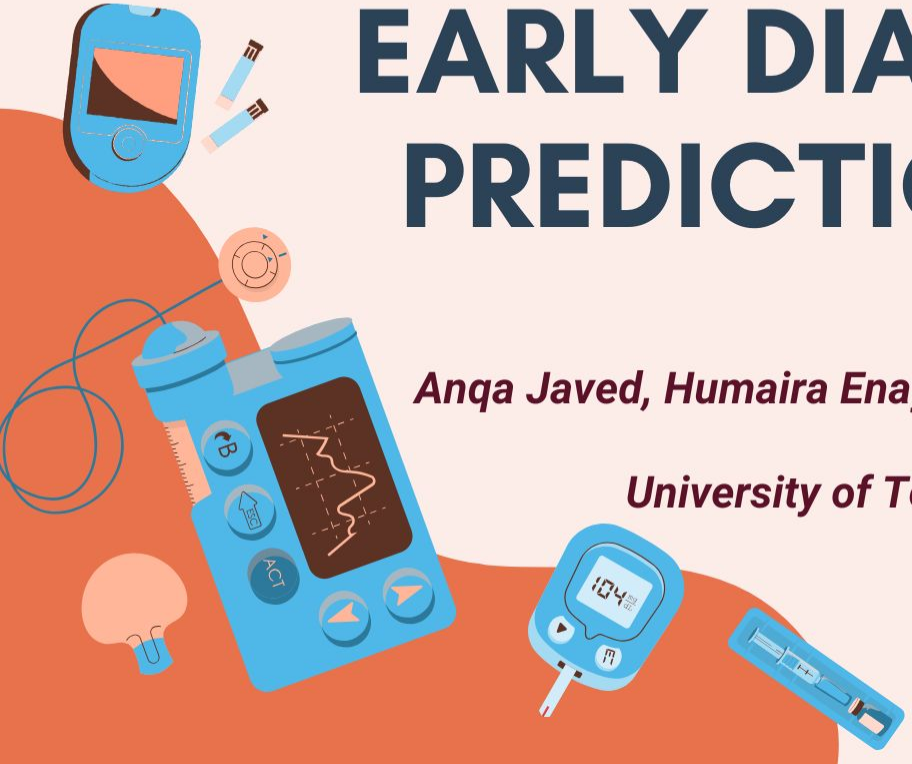
EARLY DIABETES RISK PREDICTION MODEL



Authors:
Anqa Javed, Humaira Enayetullah, Julio Carneiro, Shahab Eshghifard

University of Toronto Data Analytics Bootcamp

June 10, 2025

Table of Content

*"Based on latest International Diabetes Federation (2025) report, **1 in 9 adults** (20-79 years) are living with diabetes, with **over 4 in 10 unaware** that they have the condition.*

*Moreover, **over 90%** of people with diabetes have **type 2 diabetes**, which is driven by socio-economic, demographic, environmental, and genetic factors."*

Project Overview

This project develops a machine learning model to predict an individual's risk of diabetes. The model aims to:

- Predict diabetes risk using CDC Behavioral Risk Factor Surveillance System health data.
- Targets early intervention for high-risk individuals.
- Deployed as an interactive web app.

Objectives

1. Develop a machine learning model to predict an individual's risk of developing diabetes based on clinical and lifestyle data

2. Analyze which health factors are most important in predicting diabetes risks.

3. Create visual insights that highlight how factors like BMI, age, blood pressure, and income correlate with diabetes risk.

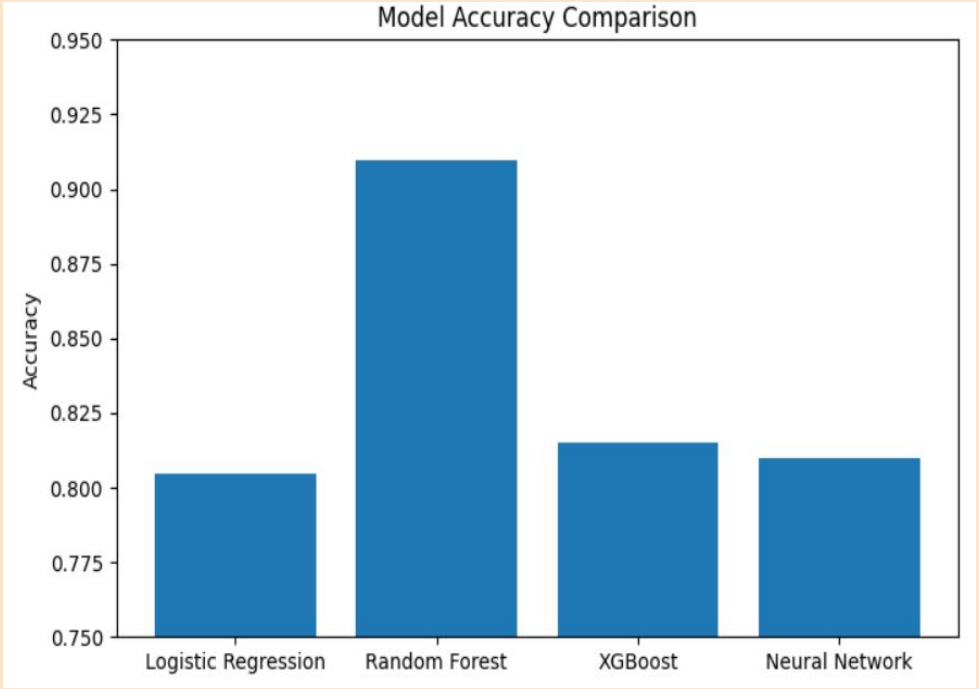
4. Deploy an interactive tool where users can input lifestyle data and receive a personalized diabetes risk score.

Tools & Technologies

- **Dataset Source:** Diabetes Health Indicators Dataset (Kaggle) with 253,680 records
- **Python:** pandas, numpy, scikit-learn, matplotlib, seaborn, XGBoost
- **Visualization:** Plotly, Tableau
- **Deployment:** Gradio
- **Development Environment:** Jupyter Notebook

Machine learning Model

- We used a **Random Forest Classifier** to predict diabetes risk.
- Random Forest builds **many decision trees** and combines their predictions.
- Each tree makes a decision based on health features like:
  BMI, Age, Blood Pressure, Physical Activity
- The final prediction is based on the **majority vote** from all trees.

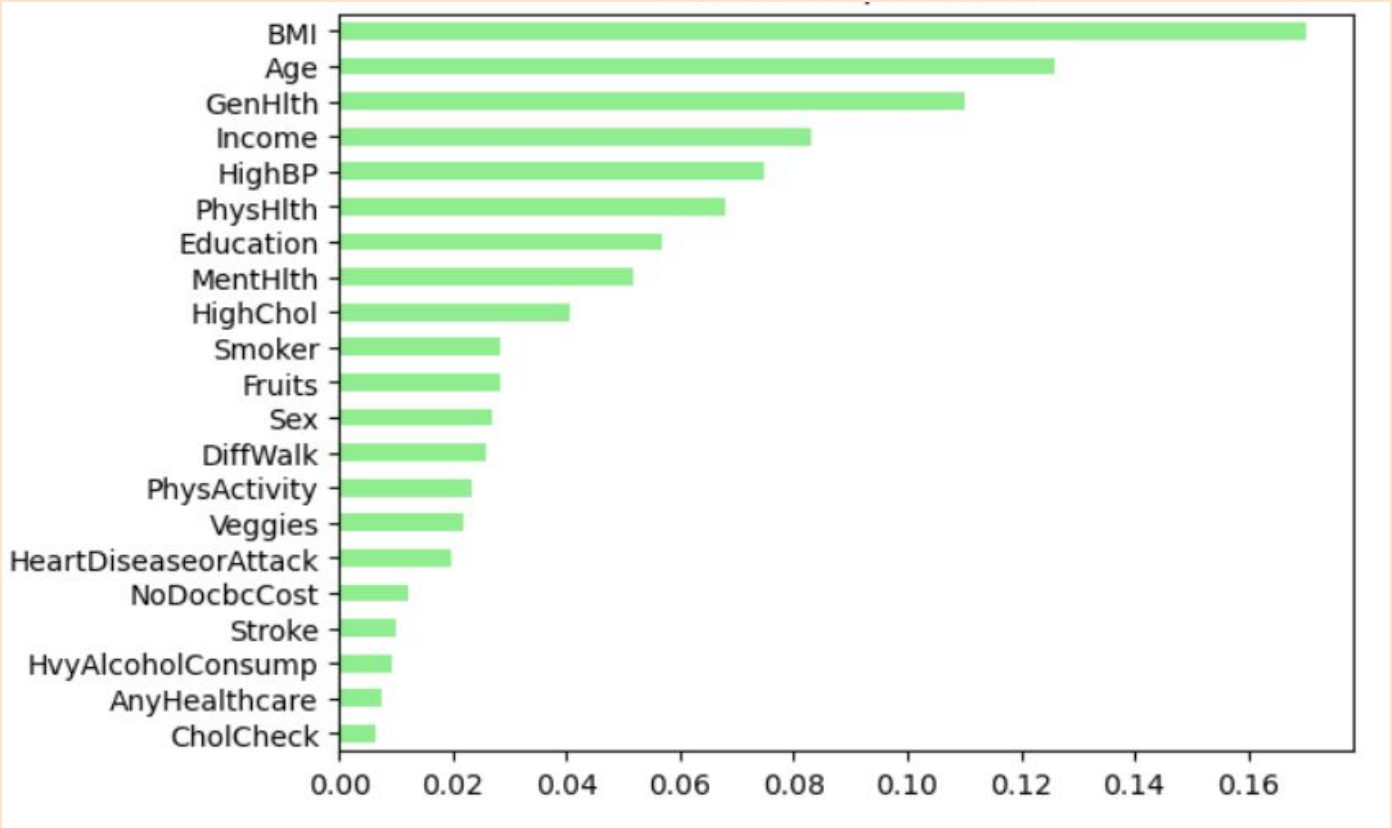


🧠 **Formula:**

$\hat{y} = \text{majority\_vote}(T_1(x), T_2(x), \ldots, T_□(x))$ Where:

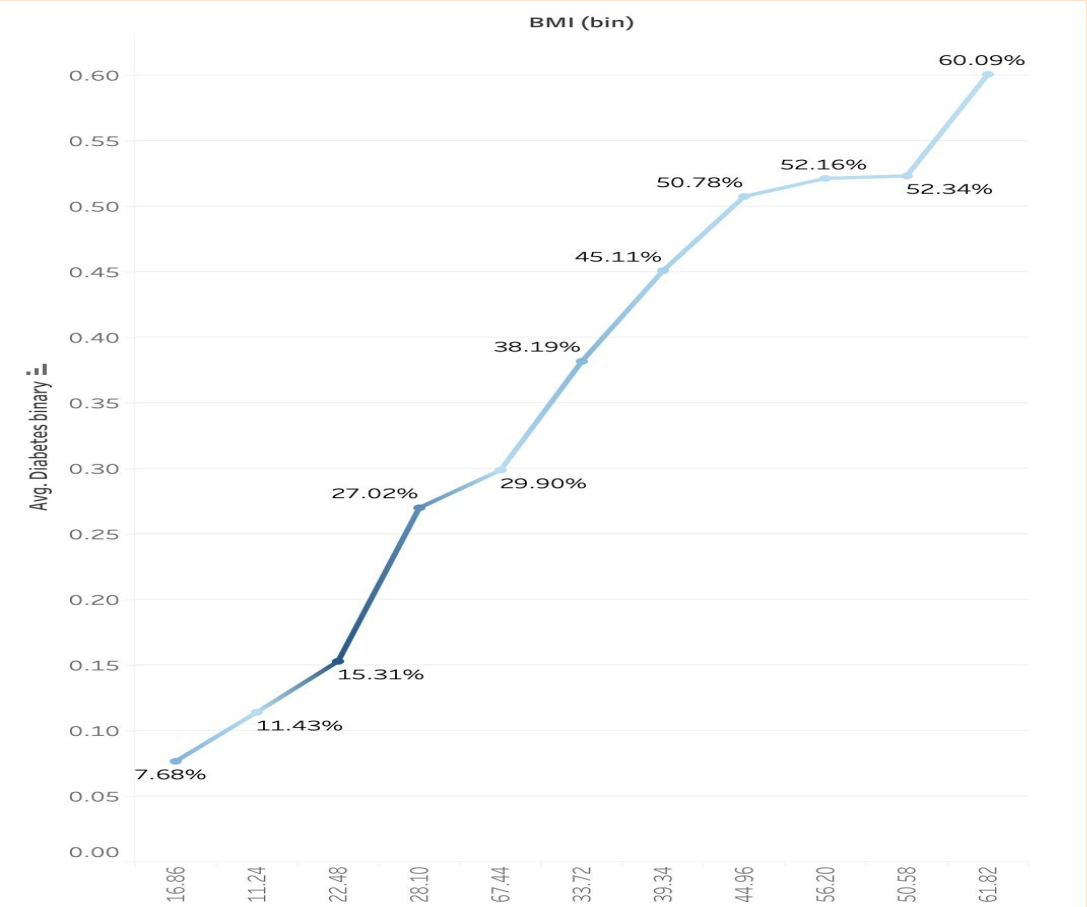
- $T(x)$ = prediction from each decision tree
- x = input features (BMI, Age, etc.)
- $\hat{y}$ = final prediction (0 = No Diabetes, 1 = Diabetic)

Feature Importance



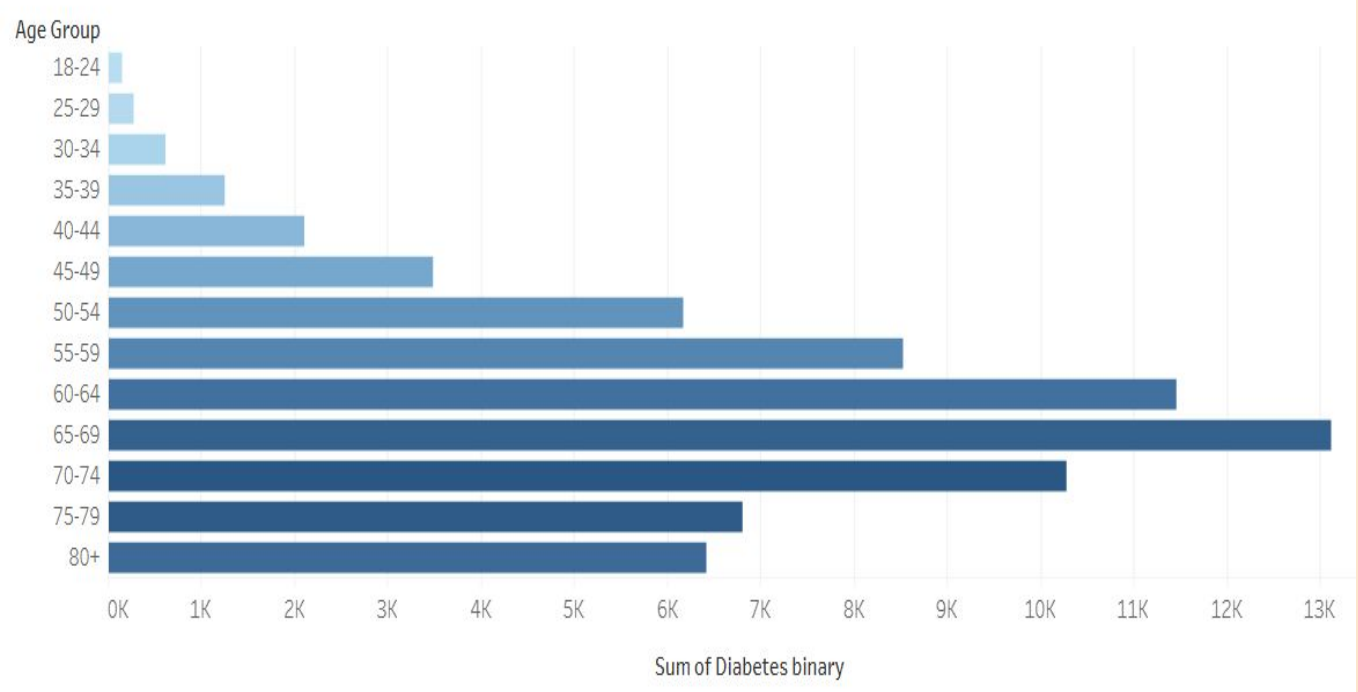
BMI and Diabetes Prevalence

- Diabetes prevalence increases with higher BMI across both males and females.



Age and Diabetes Prevalence

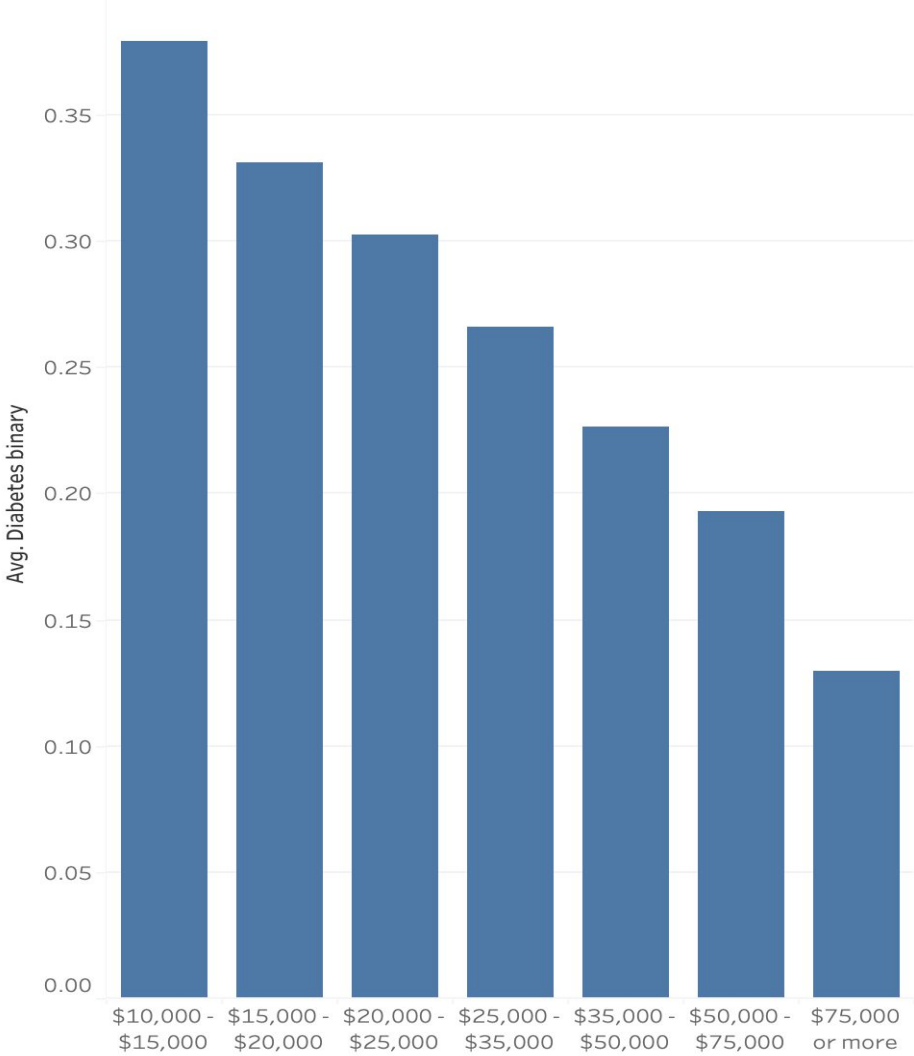
- Diabetes prevalence increases steadily with age.
- Individuals over 60 show significantly higher risk compared to younger groups.
- This trend holds for both sexes.



Objective 3

Income and Diabetes Prevalence

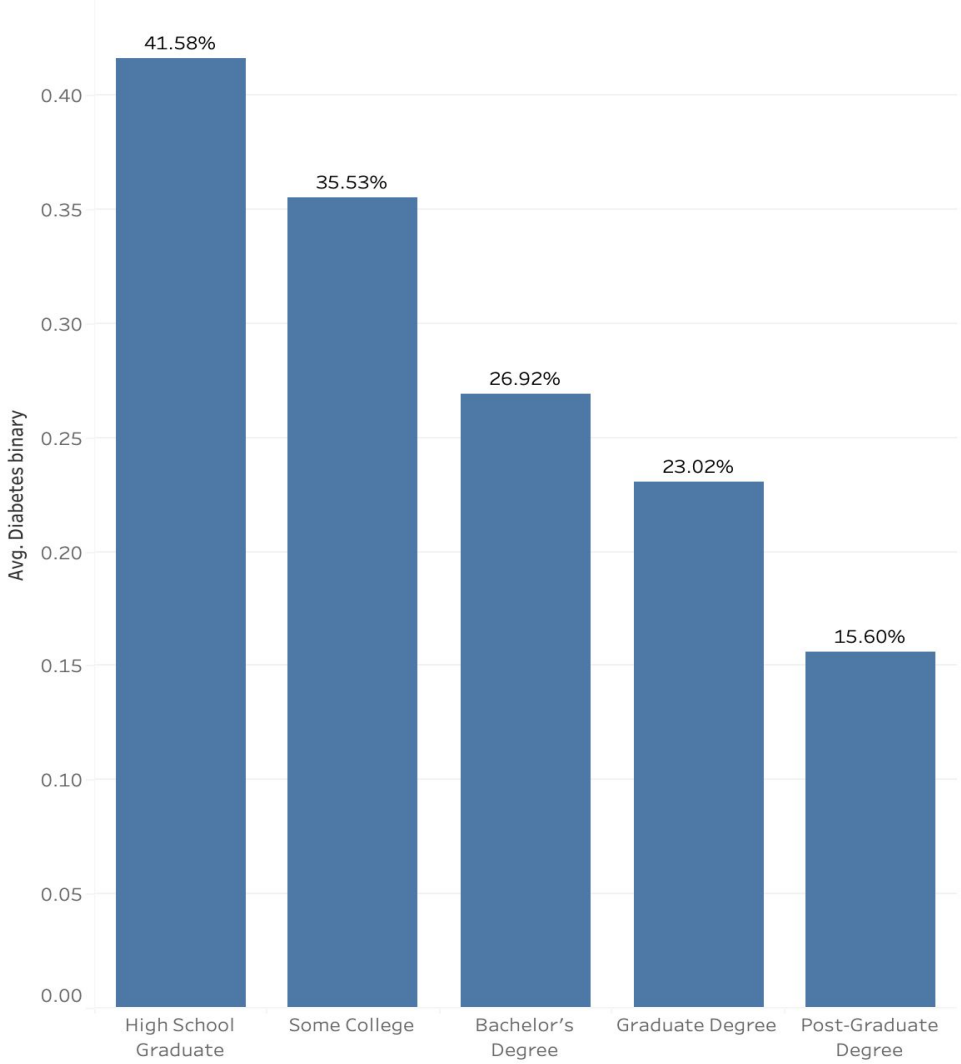
- Diabetes prevalence decreases as income increases.
- Individuals in lower-income brackets show higher risk, highlighting the need for targeted interventions.



Objective 3

Education and Diabetes Prevalence

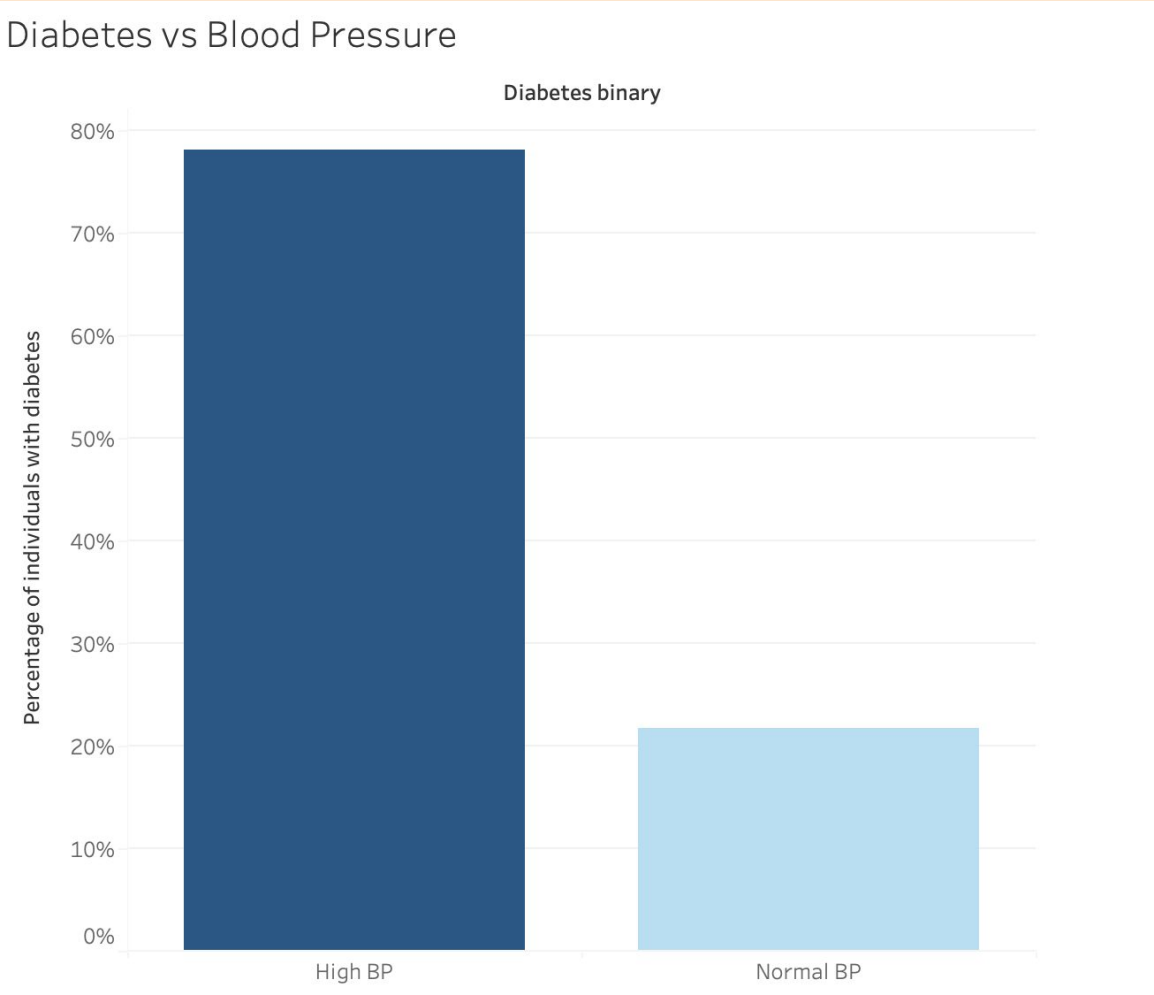
- Higher level of education influence diabetes prevalence.



Objective 3

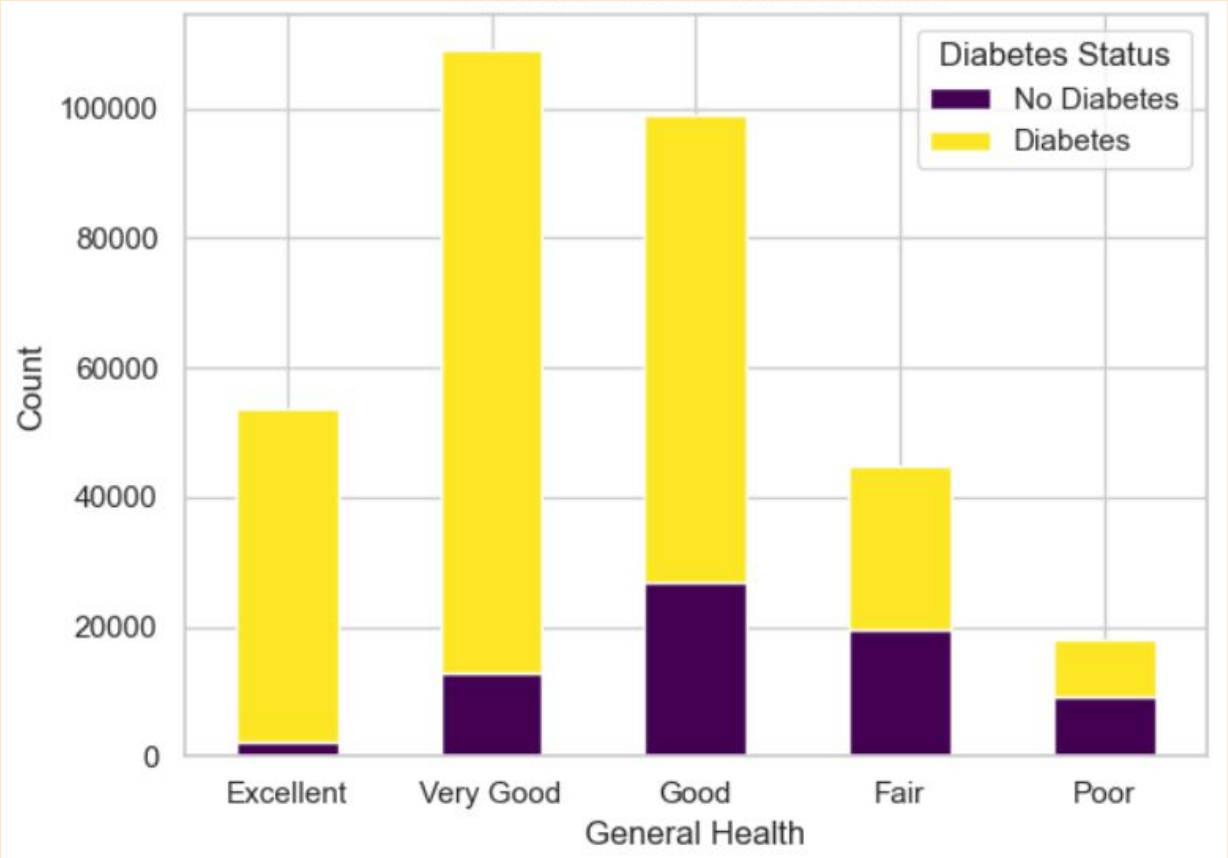
High blood pressure and Diabetes Prevalence

- Diabetes is more common among those with high blood pressure
- This happens because High blood pressure and diabetes share same risk factors: obesity, inactivity, poor diet, age, genetics.

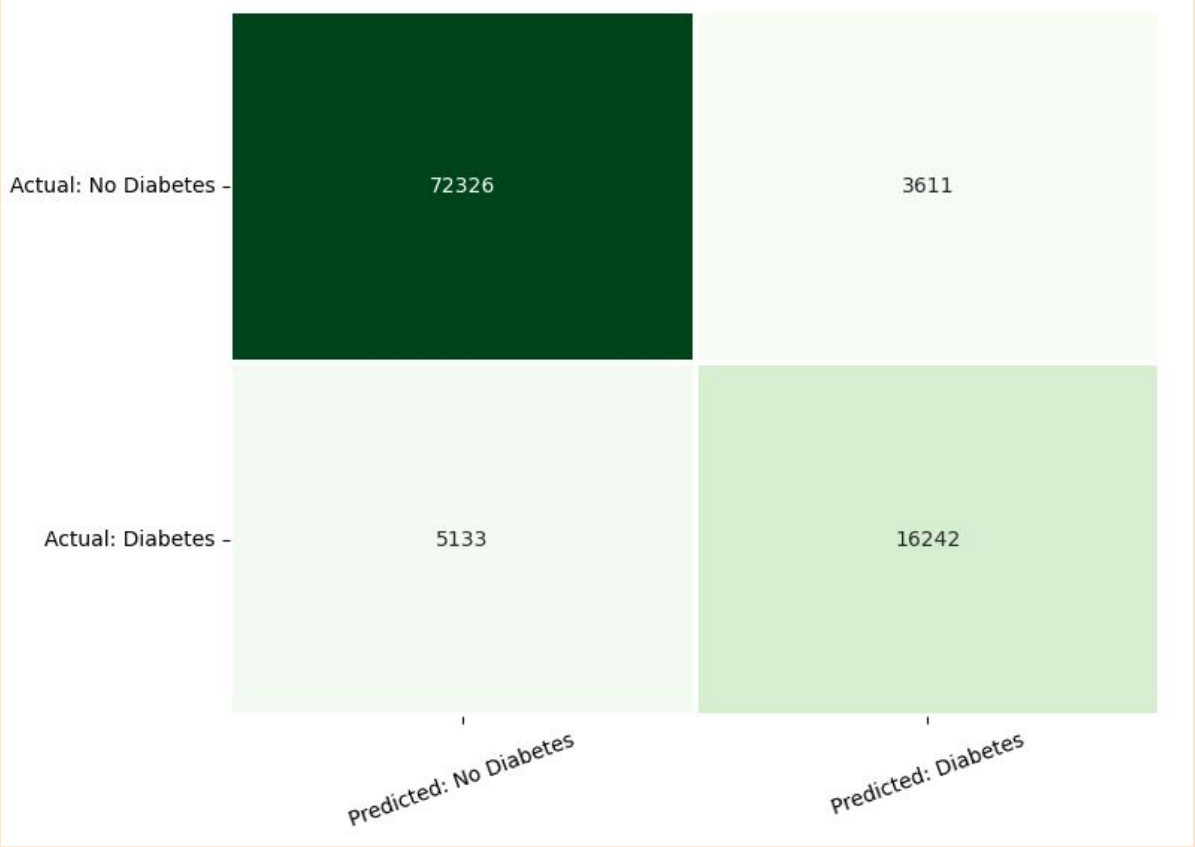


General Health and Diabetes Prevalence

- “Fair” and “Poor” categories have a higher proportion of people with diabetes
- The absolute count of people with diabetes is high in the “Good” and “Very Good” categories due to the larger population size in those categories



Confusion Matrix

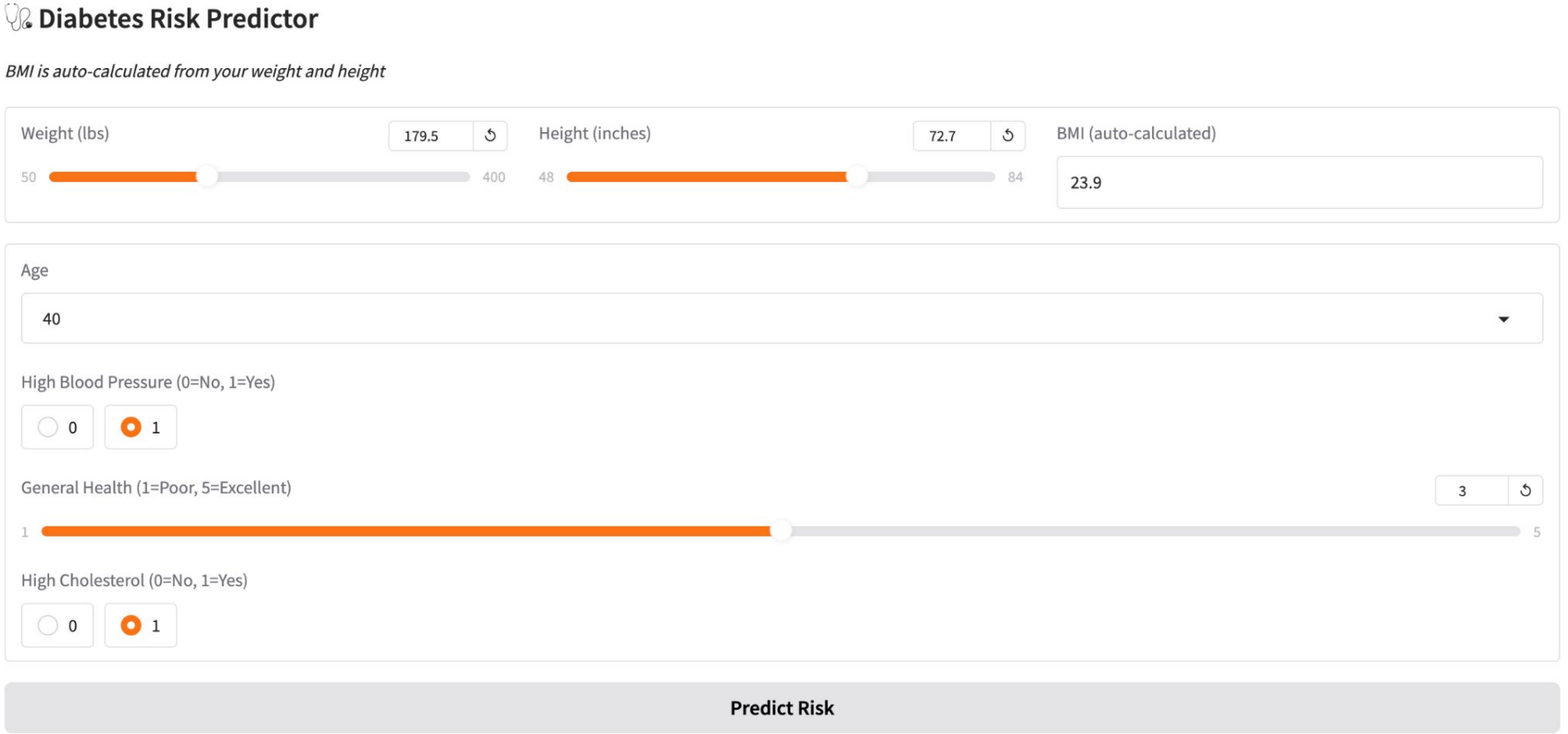


Classification Report

Classification Report:

	precision	recall	f1-score	support
0.0	0.93	0.95	0.94	75937
1.0	0.82	0.76	0.79	21375
accuracy			0.91	97312
macro avg	0.88	0.86	0.87	97312
weighted avg	0.91	0.91	0.91	97312

- Model has **91% accuracy** in predicting diabetes



Access Live Demo

Conclusion and Limitations

Conclusions

- Developed Random Forest classifier with 91% accuracy using clinical data.
- Identified top risk factors: BMI, age, blood pressure, and income.
- Actionable Prevention Tool: Interactive app provides immediate risk scores to empower users and supports early detection to promotes proactive health management.

Limitations

- The model lacks genetic history (known to affect diabetes risk).
- Data used for training may not fully represent all demographic groups.

Next Steps

- Add dietary patterns and medication history inputs.
- Develop risk stratification (Low/Medium/High) instead of binary output.

References

Dataset:

- Behavioral Risk Factor Surveillance System (BRFSS). Centers for Disease Control and Prevention (CDC). (2015). Diabetes Health Indicators Dataset. Retrieved from Kaggle: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/data

Tools:

- **Scikit-learn (sklearn):** https://scikit-learn.org/stable/. *Documentation for models, metrics, and preprocessing*.
- **Pandas:** https://pandas.pydata.org/. Data manipulation and analysis.
- **NumPy:** https://numpy.org/. Numerical computing in Python.
- ChatGPT

Tableau Live Link:

- https://public.tableau.com/app/profile/shahab.eshghifard/viz/Book1_17491728312230/ExploringDiabetesRiskFactors

THANK YOU